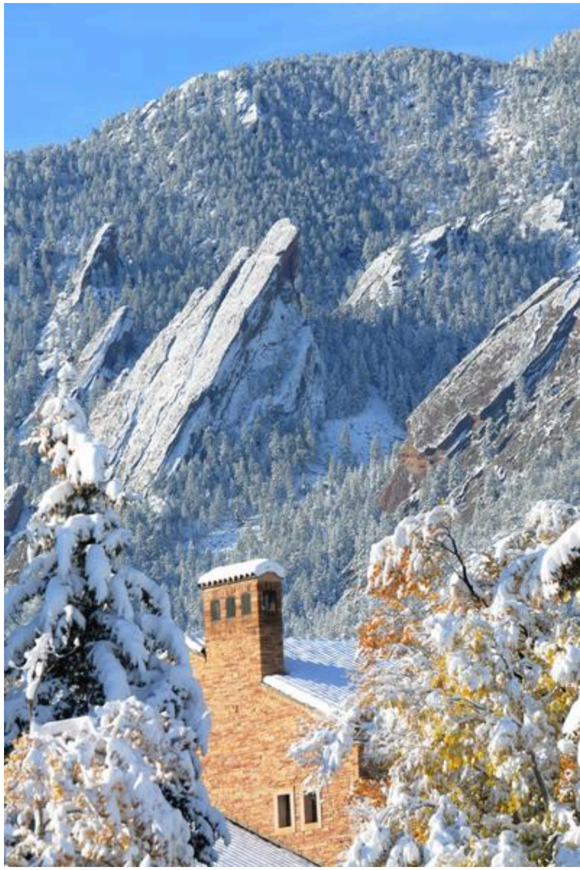
University of Colorado
Boulder

Lecture 3: Genome-Wide Analysis

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado



Harnessing the Information

OVERVIEW

- Study Design
- Analytic Challenges
- Analytic Considerations
- Population Stratification

Study Design

Study Design

- Often neglected in genetic research
  - See population stratification (later)
- The most popular design has been case-control studies
- However, cohort studies and family studies serve an important role

Case Control

- Dichotomous outcome
- Efficient for diseases of low prevalence
- Control selection *very* important
- Often nested within larger cohort study
- Examples
  - WTCCC
  - Psychiatric Genetics Consortium

Cohort Study

- Ideal for more common diseases/disorders
- Quantitative, discrete/binary traits
- Examples
  - Framingham Heart Study (FHS)
  - Agincourt

Family-Based

- Covered later

Analytic Challenges

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing Multiple Testing Multiple Testing

Multiple Testing

- GWAS
  - 1 phenotype
  - 1,000,000 markers
    - ~50,000 p-values < 0.05
- Whole Genome Sequencing
  - 1 phenotype
  - 3B base pairs
    - ?????

Addressing Multiple Testing

- Family-Wise Error Rate (FWER)
  - Bonferroni
- False Discovery Rate (FDR)
  - and variations of…
- Bayesian Approaches
  - and variations of…
- Weighted Hypothesis Testing

Dealing with Multiple Testing

- Brute Force approach comes at a cost
  - Very large samples (time/effort/resources)

- We are inherently limited in what we will be able to uncover using traditional statistical methods

GWAS to Generate Hypotheses

- No one will (or should) take a GWAS finding at face value
  - Replicate
  - Replicate
  - Replicate
- Many journals don't accept association findings without independent replication

Analytic Considerations

Coding Genotypes

- Assume a biallelic marker (SNP)
- There are three possible genotypes
  - AA
  - Aa
  - aa

Coding Genotypes

	Genotype		
	aa	*aA*	*AA*
Genotype (A)	0,0,1	0,1,0	1,0,0
Additive (A)	0	1	2
Dominant (A)	0	1	1
Recessive (A)	0	0	1

Genotype Coding

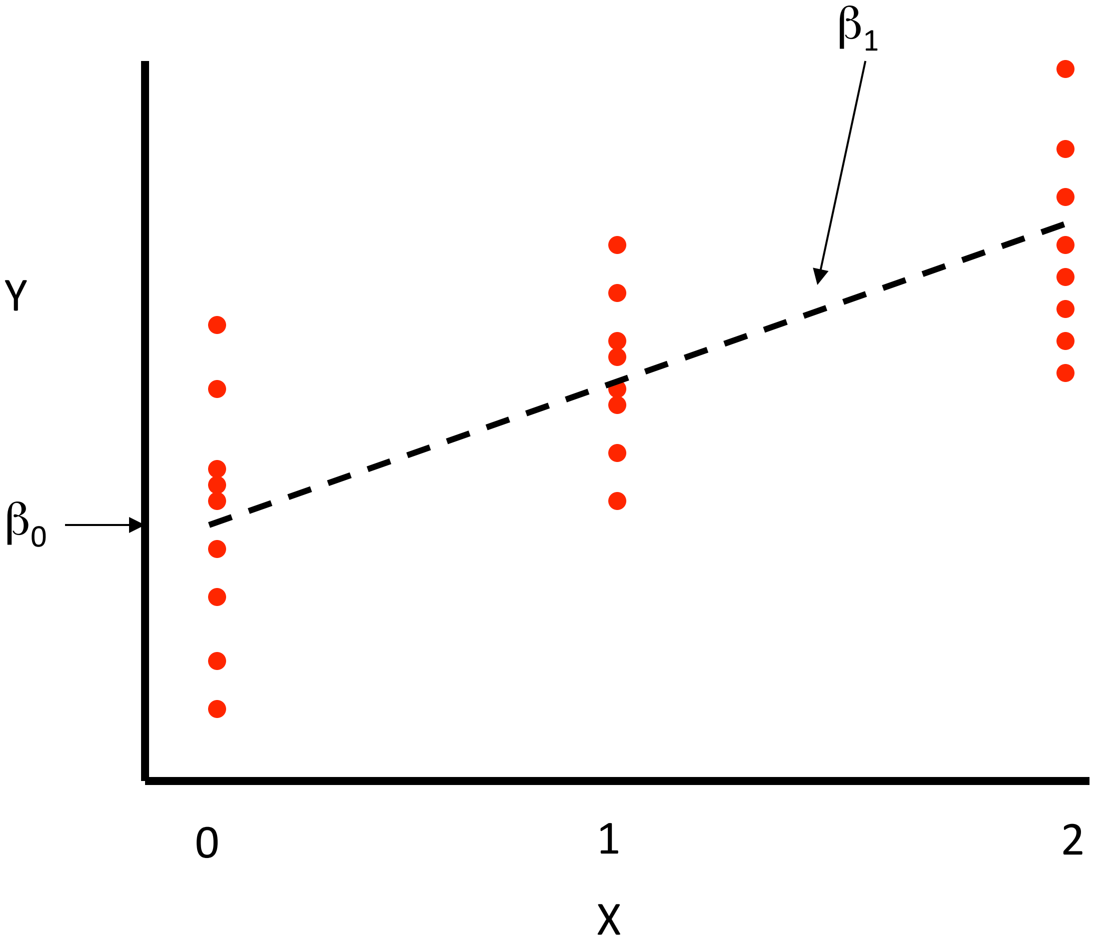
$$\text{Marker Score} = X$$

$$\text{Additive}: X = (0, 1 \text{ or } 2)$$

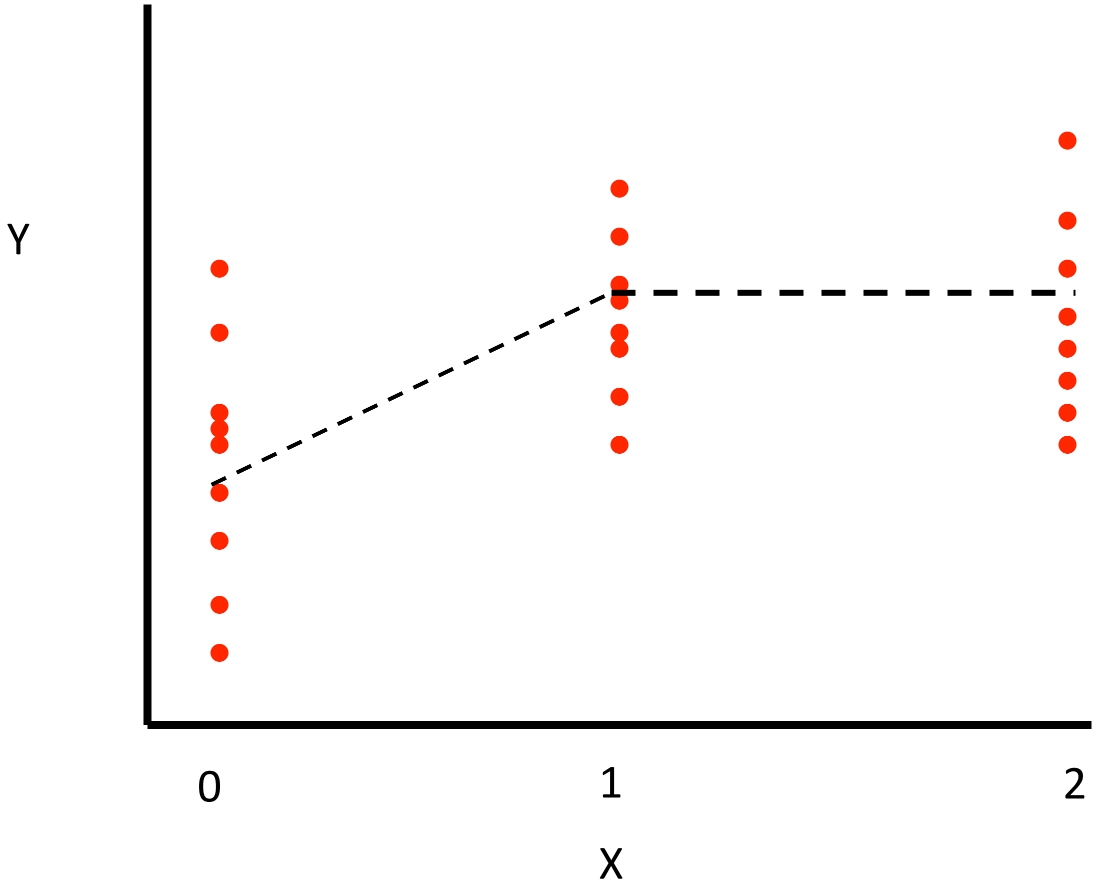
$$\text{Dominant}: X = (0 \text{ or } 1)$$

$$\text{Recessive}: X = (0 \text{ or } 1)$$

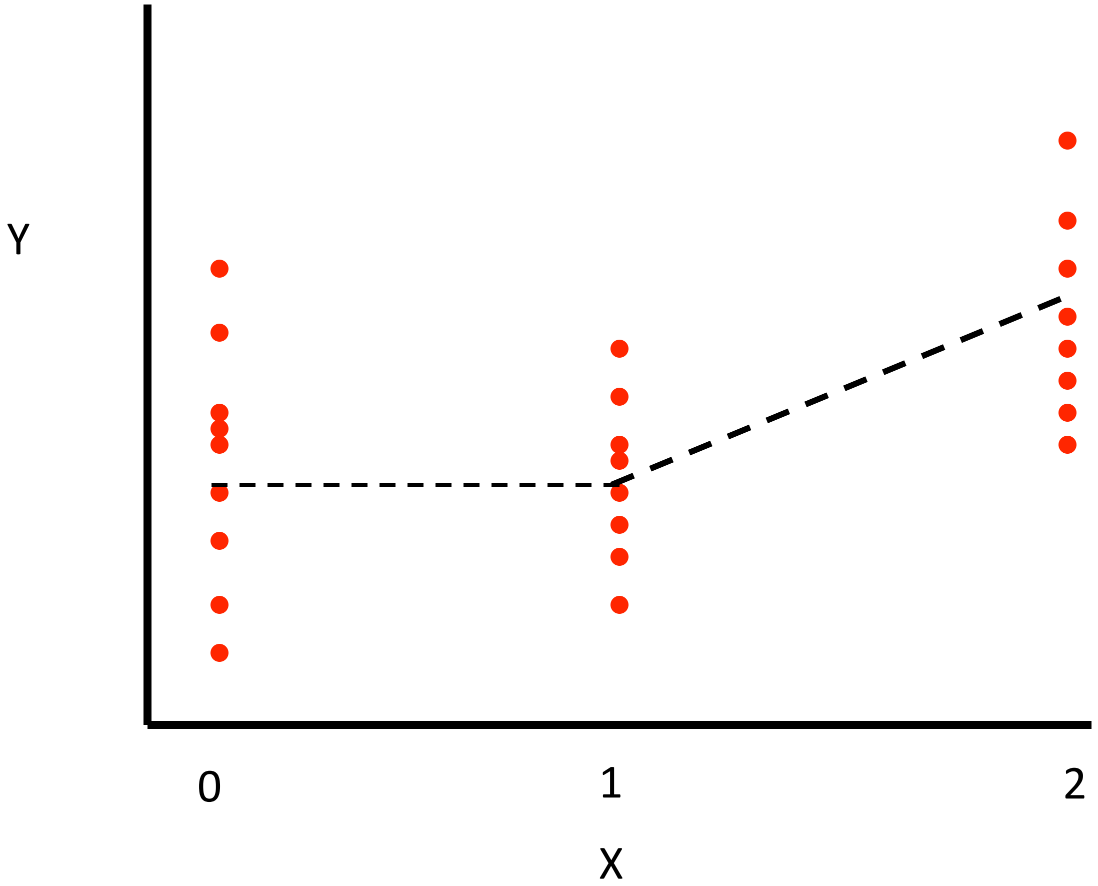
Additive Mode



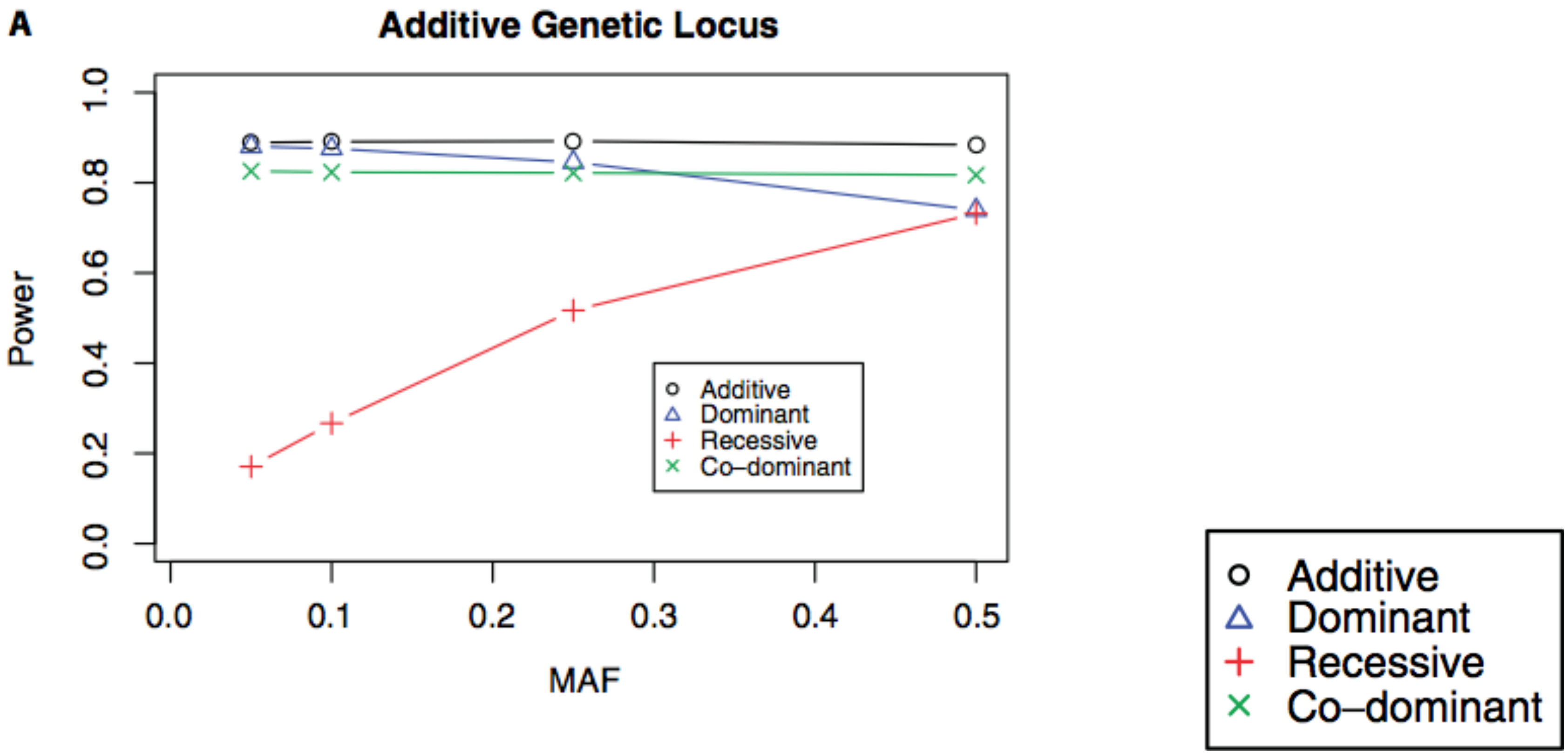
Dominant Mode



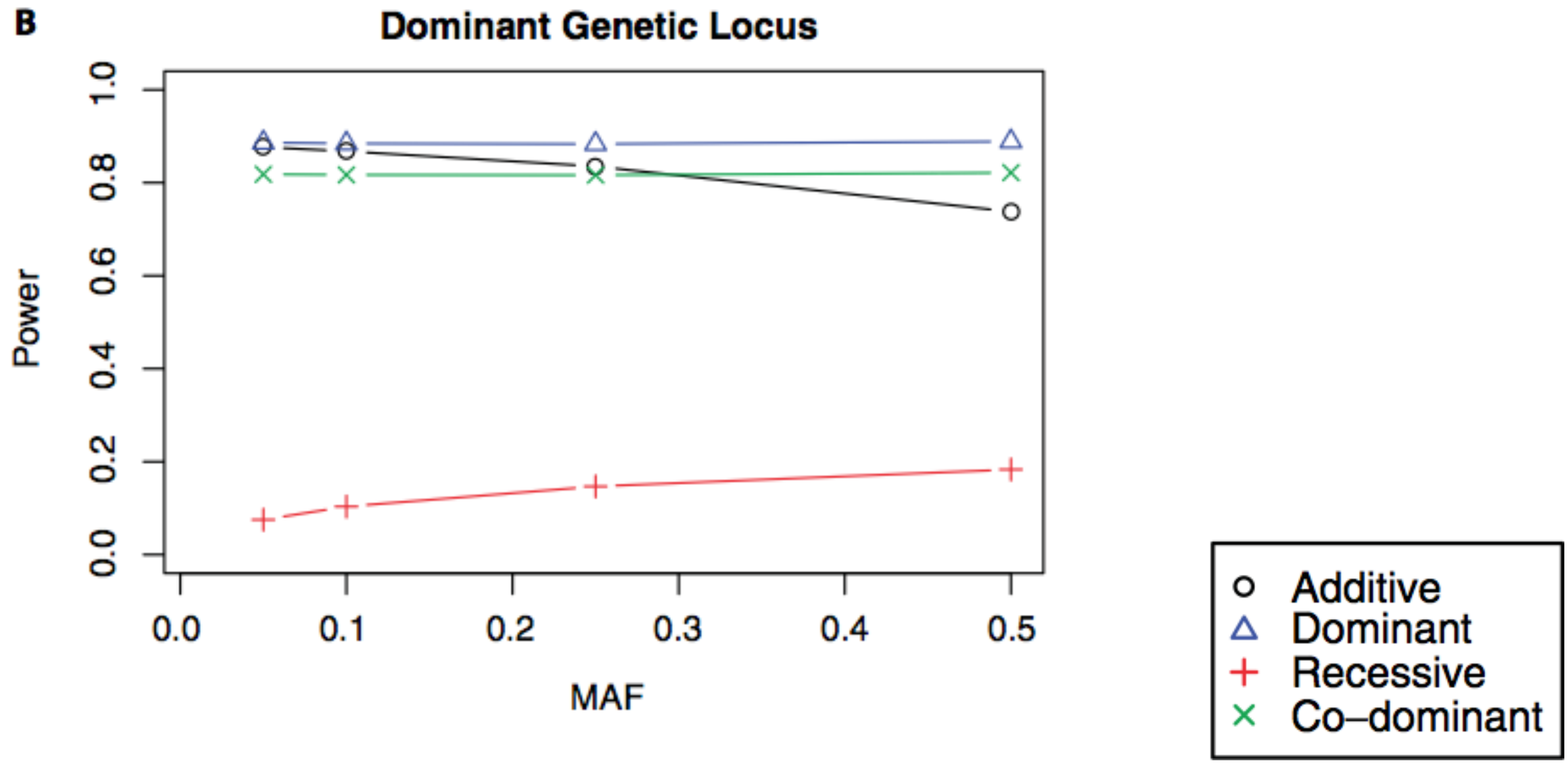
Recessive Mode



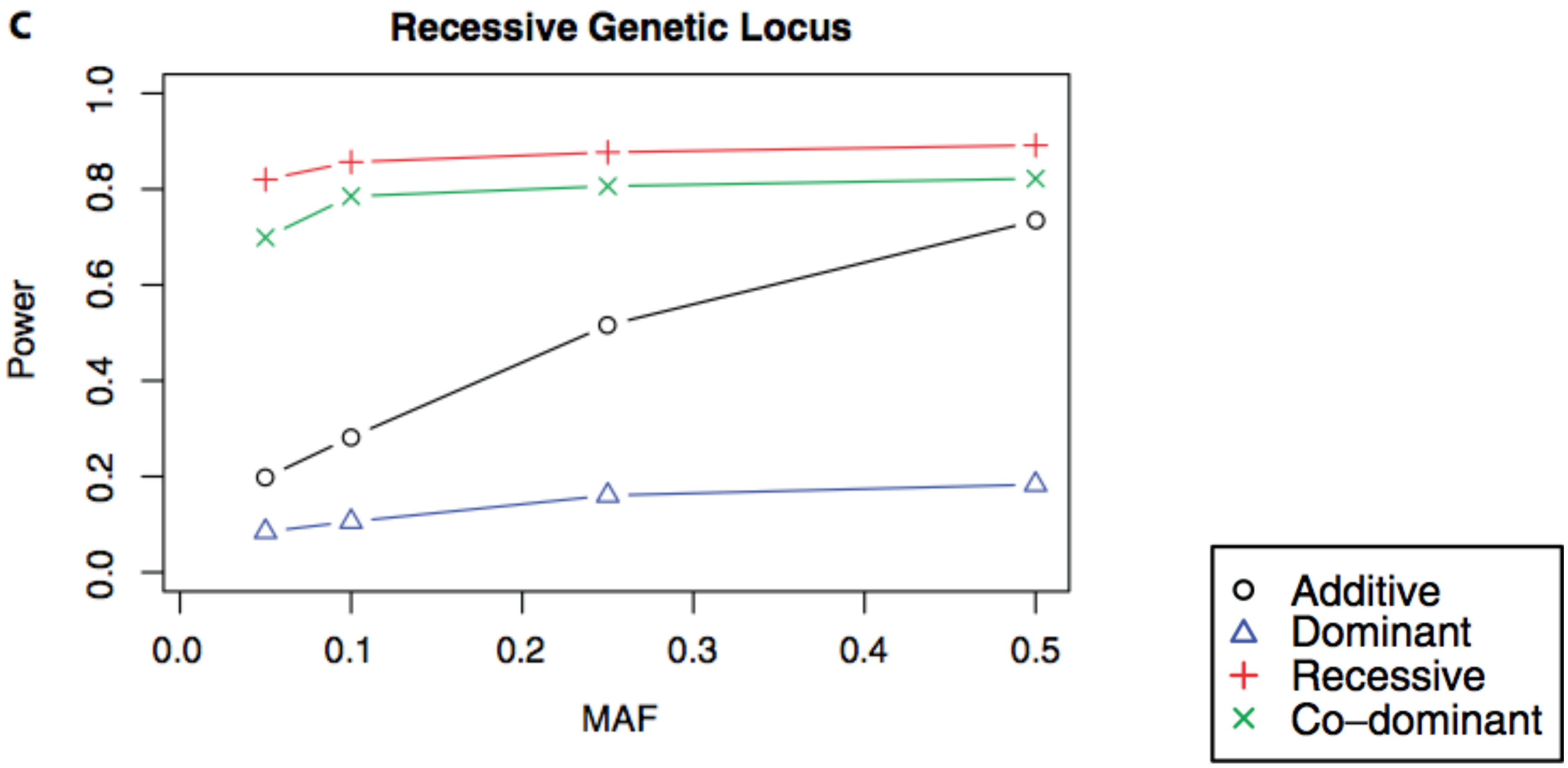
Genotype Coding…



Genotype Coding…



Genotype Coding…



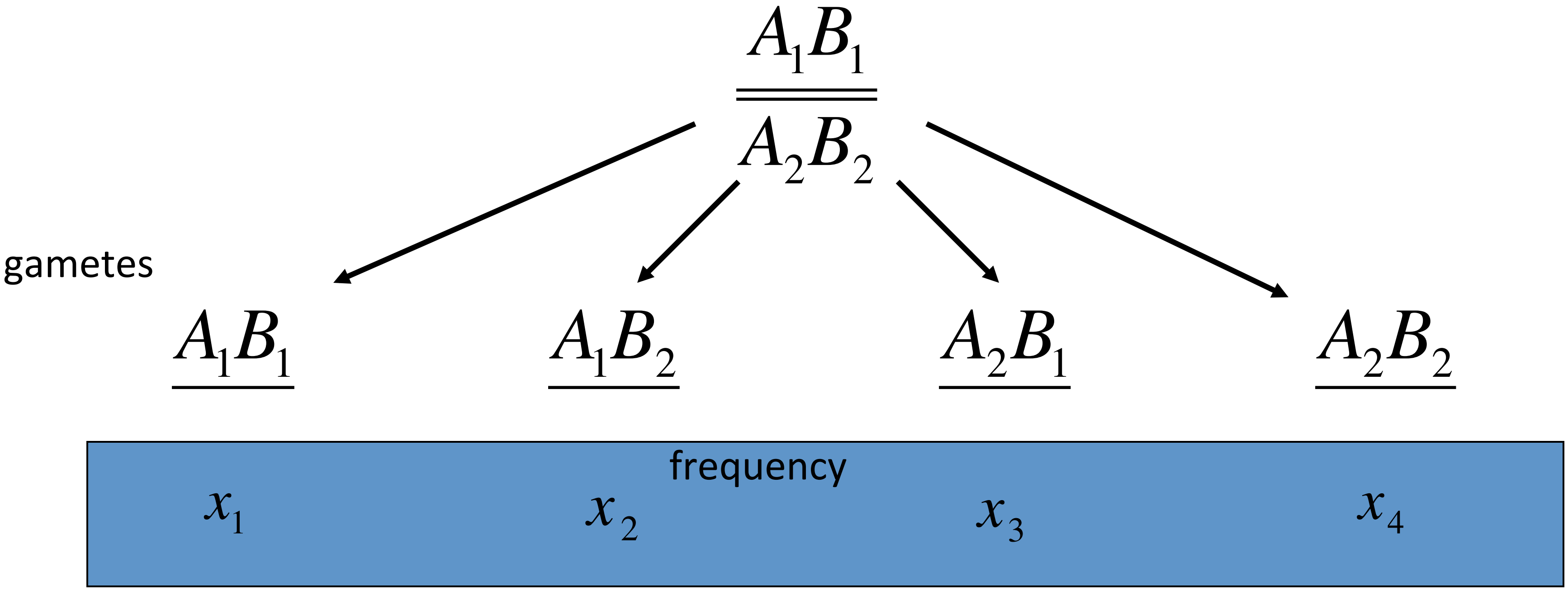
What types of analyses?

- Anything goes!
  - Typically, one large “do loop”
- Dichotomous phenotypes
- Quantitative phenotypes
- Time to onset
- Cross-sectional, longitudinal

Limitations

- Software
  - PLINK will only take you far
  - May need to write custom scripts to get what you want
    - SAS, R, SPSS, STATA, etc

Linkage Disequilibrium



$$\frac{A_1B_1}{A_2B_2}$$

gametes

$\underline{A_1B_1}$ $\underline{A_1B_2}$ $\underline{A_2B_1}$ $\underline{A_2B_2}$

frequency

x_1 x_2 x_3 x_4

Linkage Disequilibrium

Gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency	x_1	x_2	x_3	x_4

Linkage Disequilibrium

Gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency	x_1	x_2	x_3	x_4

Allele	A_1	A_2	B_1	B_2
Frequency	$p_{A1}=x_1+x_2$	$p_{A2}=x_3+x_4$	$p_{B1}=x_1+x_3$	$p_{B2}=x_2+x_4$

Linkage Disequilibrium

Gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency	x_1	x_2	x_3	x_4

Allele	A_1	A_2	B_1	B_2
Frequency	$p_{A1}=x_1+x_2$	$p_{A2}=x_3+x_4$	$p_{B1}=x_1+x_3$	$p_{B2}=x_2+x_4$

D = Observed - Expected

$$D = x_1 - p_{A1}p_{B1}$$

$$D = x_1 - (x_1 + x_2)(x_1 + x_3)$$

$$\boxed{D = x_1x_4 - x_2x_3}$$

Linkage Disequilibrium

After one generation of random mating:

$$x_1' = x_1 - \theta D$$

$$x_2' = x_2 - \theta D$$

$$x_3' = x_3 - \theta D$$

$$x_4' = x_4 - \theta D$$

$$D_{t=1} = x_1'x_4' - x_2'x_3'$$

$$D_{t=1} = (1-\theta)D$$

After *t* generations:

$$D_t = (1-\theta)^t D_0$$

What does this mean?

$$D_t = (1-\theta)^t D_0$$

D_0	*theta*	*t*	*D*
1	*0.5*	*10*	*0.001*
1	*0.1*	*10*	*0.35*

Normalized LD Parameters

$$\boxed{D' = \frac{D}{D_{\max}}}$$

D_{max} $= min(p_{A1}p_{B2},p_{A2}p_{B1})$ *if D is positive*

$= min(p_{A1}p_{B1},p_{A2}p_{B2})$ *if D is negative*

Now, LD ranges from -1 to +1

r^2

Another commonly used LD measure

$$\boxed{r^2 = \frac{D^2}{p_{A1}p_{A2}p_{B1}p_{B2}}}$$

Reasons for LD

- Mutation
- Population Subdivision
- Genetic Drift
- **Lack of Recombination**
- Selection
- Non-random Mating

LD in GWAS

- SNP markers that are in close proximity may be picking up the same signal
- One typically sees a cluster of significant p-values around a signal
- Two SNPs associated

Population Stratification

Genetic Associations

- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Population stratification

Stratification

- Essentially a confounder!

- How does it happen?

How Does it Happen?

- Two Necessary Components:
  - Subpopulation 1 has higher prevalence (mean) of disease
  - Subpopulation 1 has different allele frequency

Examine the Data

- Allele frequencies in ethnic subgroups
- Prevalence (means) in ethnic subgroups

Famous Example

Knowler et al (1988)

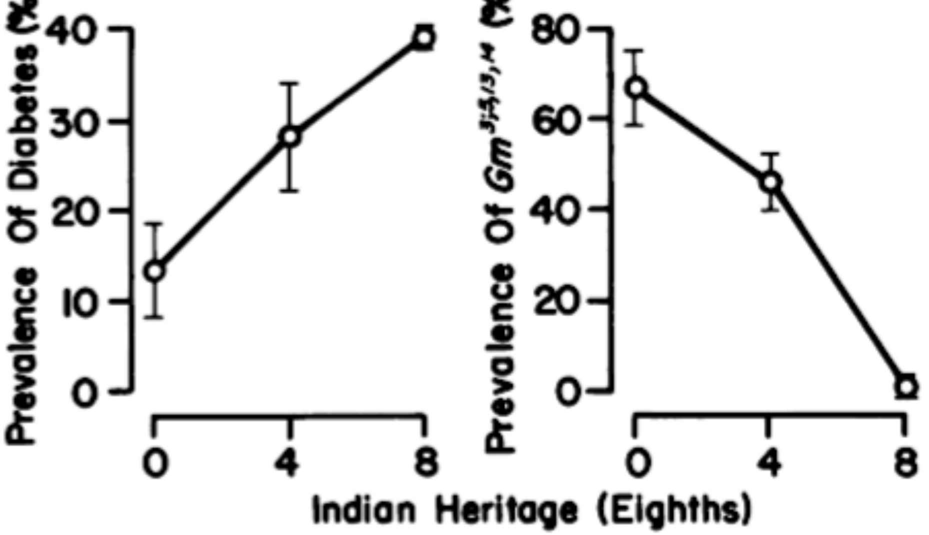
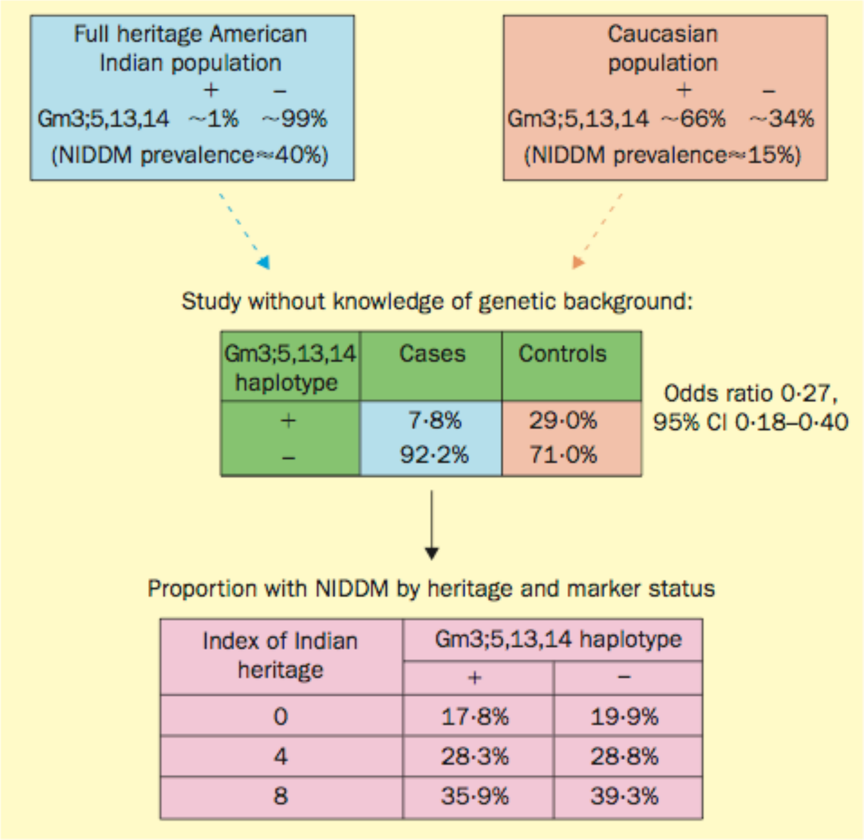


Figure 3 Age-adjusted prevalence (±1 standard error) of diabetes (left) and of $Gm^{3;5,13,14}$ (right), according to Indian heritage, among residents of the Gila River Indian Community.

Cardon et al (2003)



Full heritage American Indian population

	+	–
Gm3;5,13,14	~1%	~99%

(NIDDM prevalence≈40%)

Caucasian population

	+	–
Gm3;5,13,14	~66%	~34%

(NIDDM prevalence≈15%)

Study without knowledge of genetic background:

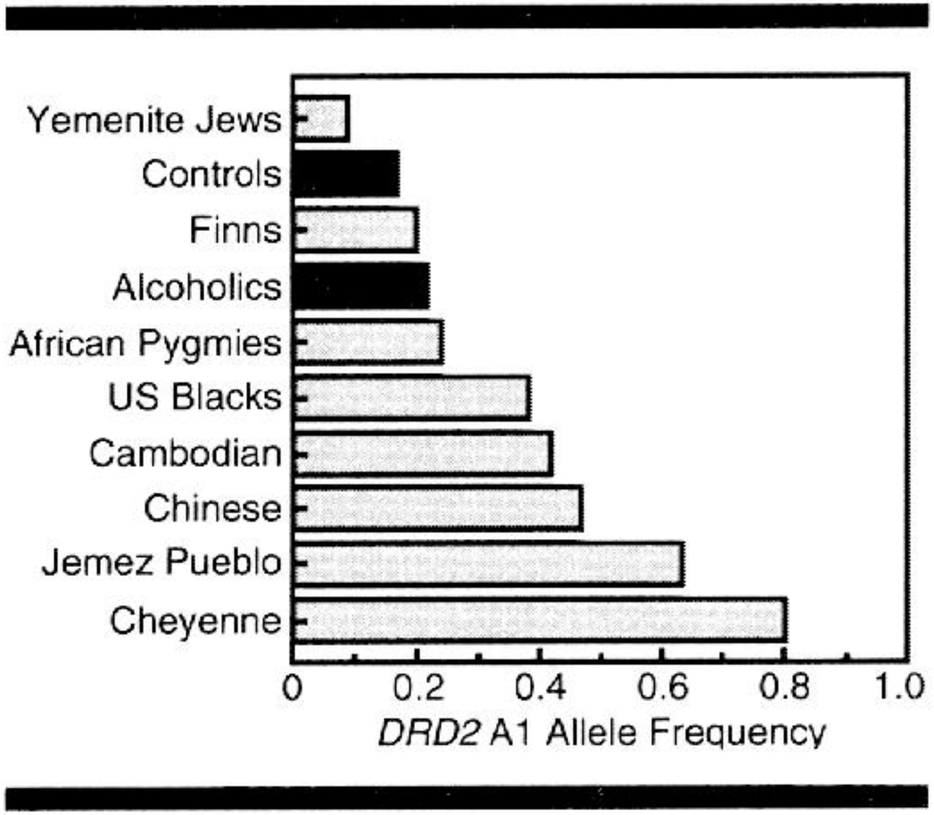
Gm3;5,13,14 haplotype	Cases	Controls
+	7·8%	29·0%
–	92·2%	71·0%

Odds ratio 0·27, 95% CI 0·18–0·40

Proportion with NIDDM by heritage and marker status

Index of Indian heritage	Gm3;5,13,14 haplotype	
	+	–
0	17·8%	19·9%
4	28·3%	28·8%
8	35·9%	39·3%

Dopamine Receptor D2



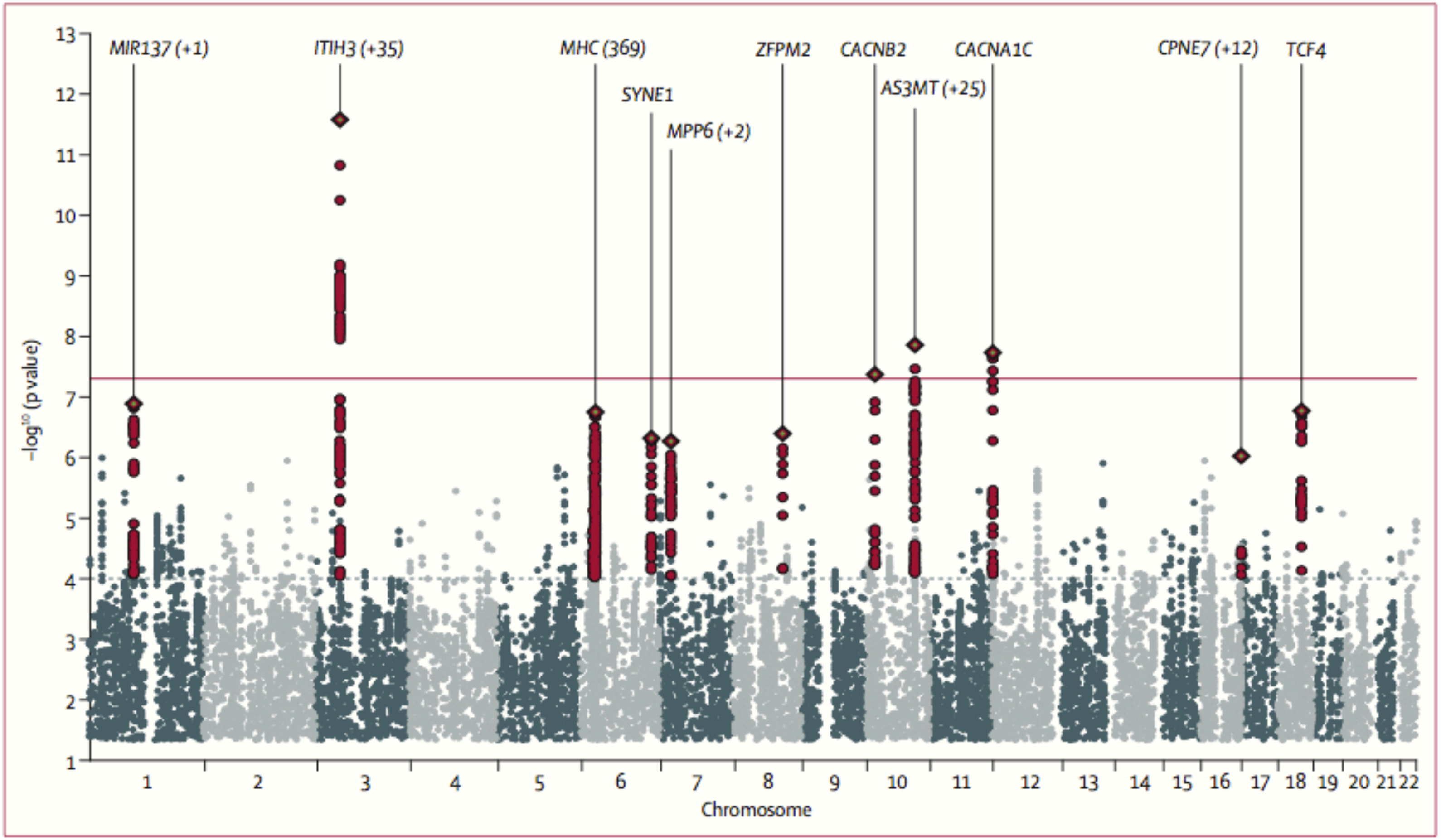
Managing Population Stratification

- Self-Reported Ancestry
  - Match (design) or Adjust (analysis)
- Use other genetic markers (ancestry informative)
  - Genomic Control (Devlin – U of Pittsburgh)
  - STRUCTURE (Pritchard – U of Chicago)
  - Eigenstrat (Reich – Broad Institute/Harvard)
  - Multi-dimensional scaling (MDS – PLINK)
- Use a family-based design

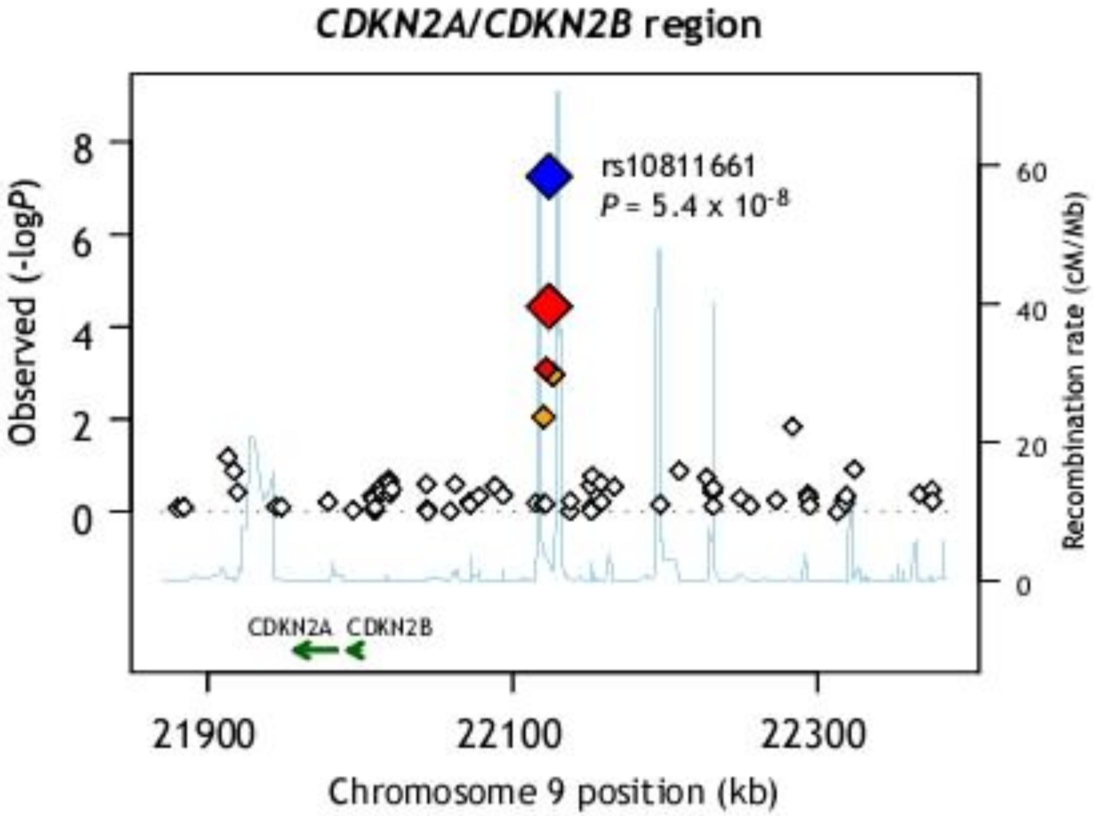
Displaying GWAS Results

- Typically, investigators will graphically display results using a Manhattan Plot
- If there is an interesting signal, investigators might also generate a regional plot
- They will also generate a quantile-quantile (QQ) plot to inspect results

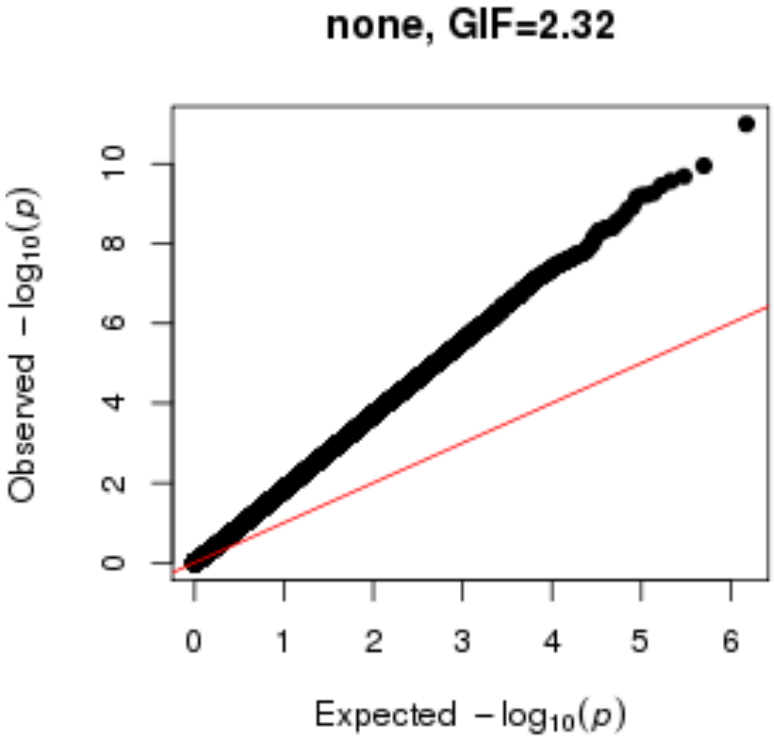
Manhattan Plot



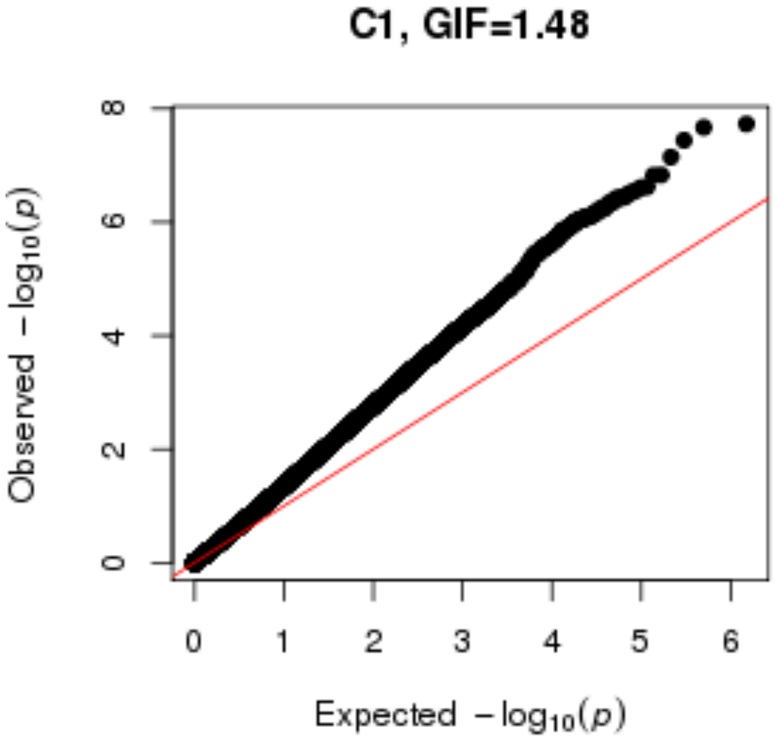
Regional Plot



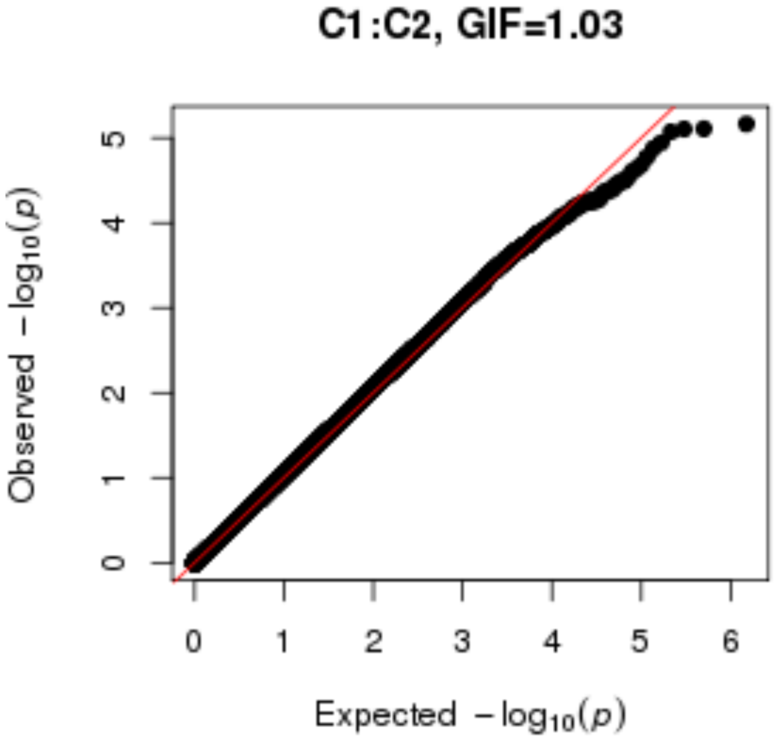
QQ Plot (unadjusted)



QQ Plot (adjusted for 1 PC)



QQ Plot (adjusted for 2 PCs)



Next up…

- Tutorial 3